AFRL-HE-WP-TR-2006-0050

AIR FORCE RESEARCH LABORATORY

A Virtual Research Partner

Jim Cowie
Felicia Guerrero

New Mexico State University
Computing Research Laboratory
Box 30001/MSC 3CRL
Las Cruces NM 88003

February 2006

Final Report for September 2004 to October 2005

20060731027

Human Effectiveness Directorate
Biosciences and Protection Division
Biomechanics Branch
2800 Q Street, Bldg 824
Wright-Patterson AFB OH 45433-7947

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Wright Site (AFRL/WS) Public Affairs Office (PAO) and is releasable to the National Technical Information Service (NTIS). It will be available to the general public, including foreign nationals.

National Technical Information Service
5285 Port Royal Road, Springfield VA 22161

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Rd., STE 0944, Ft Belvoir VA 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2006- 0050

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

FOR THE DIRECTOR

//signed//

MARK M. HOFFMAN
Deputy Chief, Biosciences and Protection Division
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (DD-MMM-YYYY) February 2006	2. REPORT TYPE Final Report	3. DATES COVERED (From – To) September 2004 to October 2005

4. TITLE AND SUBTITLE A Virtual Research Partner	5a. CONTRACT NUMBER FA8650-04-1-6534
	5b. GRANT NUMBER N/A
	5c. PROGRAM ELEMENT NUMBER 62202F
6. AUTHOR(S) Jim Cowie; Felicia Guerrero	5d. PROJECT NUMBER 7184
	5e. TASK NUMBER 02
	5f. WORKUNIT NUMBER 14

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) New Mexico State University Computing Research Laboratory Box 30001/MSC 3CRL Las Cruces NM 88003-8001	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Materiel Command Air Force Research Laboratory, Human Effectiveness Directorate Biosciences & Protection Division Biomechanics Branch Wright-Patterson AFB OH 45433-7947	10. SPONSOR / MONITOR' S ACRONYM AFRL/HEPA
	11. SPONSOR/MONITOR'S REPO RT NUMBER(S) AFRL-HE-WP-TR-2006-0050

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
Cleared as AFRL/WS-06-1489, 13 Jun 06.

14. ABSTRACT
This report describes a design study carried out at the Computing Research Laboratory, New Mexico State University. The goal was to investigate the feasibility of creating a software agent that would be able to interact with researchers and provide them with support at a level equivalent to a human research partner. This agent is referred to in the report as the Virtual Research Partner (VRP). The authors investigated what current researchers do and what tools they use. This work was based on a small set of interviews with active researchers at NMSU. The authors searched the business and research literature for equivalent projects that provide support for managers or other subject matter experts. The authors looked at research projects working on developing computers which act as assistants. We itemized the capabilities of an ideal autonomous VRP and attempted to define a feasible subset of support capabilities that would make initial versions of a VRP useful to a researcher. The authors then proposed a design based on using an ontology (world knowledge representation) and a software architecture based on blackboard systems as a basis for an extensible VRP. The actual implementation of a large coverage VRP with capabilities that support much of the research process is a massive task. One possibility is to bootstrap this effort by providing an initial system and developing a community of users to implement and improve the many capabilities needed by such a system.

15. SUBJECT TERMS
Computer Aided Research, Research Partners, Informatics Ontology, Blackboard Systems, Machine Intelligence Knowledge Acquisition, Representation, Manipulation and Management

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON: Ted Knox
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	40	19b. TELEPHONE NUMBER *(Include area code)*

PAGE IS INTENTIALLY LEFT BLANK

Table of Contents

THIS PAGE IS INTENTIONALLY LEFT BLANK

PREFACE

This report describes an investigation carried out at the Computing Research Laboratory (CRL), New Mexico State University into the possibility of developing a "Virtual Research Partner." This was a fourteen-month study, which investigated the nature of research and the current tools available to and used by researchers. The research was carried out by Ms. Felicia Guerrero, a junior in the Mechanical Engineering Department under the supervision of Dr. Jim Cowie (CRL). The work was supported by the Air Force Research Laboratory under its HBCU/MI research program (FA8650-04-1-6534). The AFRL technical representative was Dr. Ted Knox.

THIS PAGE IS INTENTIONALLY LEFT BLANK

INTRODUCTION

A Virtual Research Partner

The original call for proposals envisaged a research assistant that is truly an intelligent partner. Paraphrasing the original call for proposals -

> *Research is all about understanding what is known about a topic, then exploring the unknown and integrating the new information into a more thorough understanding of the topic . .... What capabilities would a software package need to autonomously research and write a grant proposal, research paper or assist in the scientific process? (Knox, 2004)*

Given our experiences of machines understanding language, the VRP does seem to be a pipe dream and the claims of strong artificial intelligence (AI) that propose generally intelligent machines to be possible are not held by many researchers nowadays. This, of course, depends on the definition of intelligent behavior adopted by any particular research group. A more conservative supportive AI that proposes methods to help humans using whatever computer power, massive data, and heuristic based software is, however, helping people solve problems. In this research we have explored both paradigms and tried to establish what the long-term possibility for a VRP might be. The initial study was to investigate what a few individual researchers at NMSU used in terms of software and other support to help their research. Interviews with researchers in three disciplines at New Mexico State University (NMSU): Computer Science, Psychology, and Mechanical Engineering, were carried out.

From these interviews, a description of the tools used by the researchers, as well as a supplementary list of tools needed to help them with problems that they encounter in their work was produced. In parallel, various tools used by businesses; tools to carry out group work, support planning and produce documentation, were investigated. Tools emerging to support intelligence analysis, research in which CRL itself is actively engaged; ranging from search engines to tools for link analysis and information discovery were also investigated. One example of a needed tool that is becoming important for rapidly advancing fields, such as genomics, is methods for exploring the literature. CRL has been working with Sandia National Laboratories

(VXInsight) on support for visualization tools. We have included this work as one possible method of moving forward in the direction of "intelligent" support.

General Plan of Work

It was planned to investigate the current state of software to support research by first interviewing active researchers in three departments at NMSU: Psychology, Chemistry, and Mechanical Engineering. Original intentions were to focus on Engineering Psychology, Electro-Chemistry, and Robotics, but finally worked on Engineering, Computer Science, and Psychology because of time constraints and the availability of researchers for discussion. All three departments have highly active researchers funded by NSF, DoD and NASA. A goal was to establish whether the needs and requirements of the different disciplines would call for different kinds of tools to be available. Certainly the specific software that is used to support research investigations will vary from one subject to another. For instance, the psychologist will be using SPSS (Statistical Package for the Social Sciences), the computer scientists will be using a variety of simulation and modeling packages, and the mechanical engineers will be using another set.

It was also assumed that there would be another set of issues which would be common across disciplines. Examples of these issues are; the need to maintain references, the need to produce research reports, papers, and web sites, the need to satisfy legal and budgetary regulations, or simply email. These tasks actually absorb a surprising amount of a researcher's time and energy. Then finally there are the tasks associated with communicating with colleagues and financial supporters and existing in a wider community.

It was then planned to develop architecture for the VRP. First was a general list of capabilities that the VRP would need. These would be based on our interviews of the researchers and investigations of current software packages.

Investigating Research Methodologies

The second dimension of the study will be the different needs of pure and applied research. As research becomes more applied, there is a greater need to specify the details of design to a larger group. This entails having methods for communicating the specifics of a design accurately and efficiently to the group. A greater need for testing and evaluation will probably also be found.

For the pure researcher many of the problems are keeping up with fast moving and evolving fields. Newly emerging visualization, document retrieval and intelligent browsing tools will have a significant part to play. The ability to access and use large-scale databases being produced by researchers and centers all over the world, for example the Biodynamic Database at AFRL (Air Force Research Lab, 2005) is also a significant factor in much research work. The emergence of standards such as scientific ontologies and representation methods geared towards reusability of data is also an important issue for modern researchers.

Evaluating Available Basic Software and Tools

According to Thomas Alva Edison "Invention is one percent inspiration and ninety-nine percent perspiration." This is true of much of the research process. If a suitable tool was available at the proposal writing stage then surely it should be possible to generate formats for reports, checkpoints for milestones and task lists for a researcher and his/her assistants and colleagues.

Generating bibliographies, producing posters for conferences, and the ubiquitous power-point presentations all absorb major amounts of effort. Budgeting, finance and purchasing also take up significant amounts of time. One question to answer is what current researchers do about this workload and do they have tools to help them with these mundane but necessary tasks. A large part of thr VRP investigation was dedicated to the capabilities and usefulness of available basic software packages and research tools. Specifically software packages which help in discovering information, and managing content and knowledge were examined. Other tools investigated were specific research tools such as the Science Citation Index, article databases, internet searches and other similar tools.

Advanced Tools for Collaboration and Investigation

One aspect of this study was the investigation of tools intended to stimulate researchers' thinking, which can help them in the process of discovery itself. The Google search engine is definitely becoming one of these tools, but more sophisticated methods for exploring knowledge must be found. CRL has been collaborating recently with Sandia National Laboratory on investigations in exploring the scientific literature with their tool VXInsight (Sandia National Laboratories, 2005). This tool allows researchers to cluster documents based on common content or common patterns of citation. The exciting outcome of the citation method is that new

fields can be seen to be emerging as new document clusters appear which lie between established fields and are linked to them. CRL's support here is in the area of content-based summarization using information extraction.

The Prospects for an Intelligent VRP

Question answering technologies have been first contemplated almost forty years ago, with Weizenbaum's Eliza (Weizenbaum, 1965) and Colby's Parry (Colby, 1973) programs. However, those programs did not have the benefit of the research and development in knowledge engineering, natural language processing, planning and reasoning that flourished since those times. The issue of generally intelligent communication with computers using natural language has been studied in a variety of environments, including such R&D thrusts as database front ends, dialog modeling and, more recently, a variety of applications connected with the internet. Evaluations have been carried out by the National Institute for Standards and Technology on question answering as a specific information retrieval task since 1999.

The main types of research that are germane to the VRP project are knowledge acquisition, representation, manipulation and management. A number of current and recent efforts have been devoted to building ontologies (CYC, Ontolingua and many others), both generic and domain-oriented. Some efforts have been stressing knowledge representation formalisms, others the types of logic that might be used to define the answer finding process. Some have concentrated on the problem of the acquisition of massive amounts of data needed for even a domain specific system (for example a system with a basic knowledge of chemistry). Efforts in this area have included support for domain experts, who are not knowledge engineers and who don't know the full structure of the knowledge base. Another approach is the merging and normalization of many different ontologies using automatic methods and standards for knowledge interchange.

The LOOM language (LOOM) is a mature example of the class of approaches which support a set of operations on declarative knowledge to support deductive query processing. These operations include forward chaining, unification, and truth maintenance technologies. The Information Sciences Institute, UCLA, also supports the complementary aspect of knowledge-bases; tools to "extend and modify knowledge bases" (EXPECT) and also tools to transfer knowledge from one format to another (Chalupsky, 2000).

Other approaches attempt to reduce the impact of knowledge assimilation (RELIANT). This was one of the core problems experienced in the expert system phase of AI. While a sophisticated system could be built with significant effort to solve some problem (e.g. computer room layout, or appendicitis diagnoses) it became clear that an equivalent amount of effort was required, even for closely related problems. So the new focus of effort becomes the replacement of highly skilled knowledge engineers with knowledge experts armed with sophisticated software. This completely ignores the problem that in the, as yet, far from understood area of knowledge acquisition the only sure path to a solution is controlled development supervised by experts in knowledge representation.

The research efforts at Stanford's Knowledge System Laboratory (KSL) are attuned to this problem and focus on the difficult problems of supporting knowledge acquirers in focusing their efforts. The KSL group is also developing tools to test the adequacy and correctness of knowledge bases. Tools of this type are an essential component of knowledge acquisition, supporting acquirers who cannot possibly know all the complexities of a knowledge base. The knowledge base acquisition environment (KBAE) developed for the OntoSem ontology at CRL include a validation checker for similar concepts and attempts to guide a knowledge acquirer in the correct placement of new concepts.

CYC (Lenat & Guha, 1990), is probably the most prominent of the large scale knowledge base systems and is one in which this lab has the most knowledge, having carried out a funded evaluation of CYC as a resource for natural language processing in 1996. Although of impressive size it was found that CYC had many asymmetries in its knowledge. The OntoSem ontology has the same weaknesses, which are to a large extent due to funding. OntoSem knows a lot about acquisitions and mergers and little about biology, and also partly due to the only partially constrained enthusiasms of information acquirers.

WHAT IS RESEARCH?

"The systematic, intensive study directed toward fuller scientific knowledge or understanding of the subject studied. Research may be classified as either basic or applied. In *basic* research the investigator is concerned primarily with gaining a fuller knowledge or

understanding of the subject under study. In *applied* research, the investigator is primarily interested in the practical use for the purpose of meeting a required need." (NSF, 1996)

As Dr. Ted Knox of the Air Force Research Lab stated, "...research is all about understanding what is known about a topic, then exploring the unknown and integrating the new information into a more thorough understanding of the topic."(Knox, 2004) Computer software packages that are available to assist research today should be able to do just that. These software packages are supposed to be helpful in managing necessary information to conduct research or business. But how helpful the capabilities are and how much more skill and capabilities these packages would need to research autonomously? Writing grant proposals, research papers, reports and being guided in the right direction are needs which recur multiple times throughout a research project. These could be autonomously completed by an "intelligent assistant" so researchers can focus, as well as dedicate more time, on the principal material of the research.

Research is a complex activity that varies in its nature from discipline to discipline. Another dimension which influences the nature of research is the range from preliminary investigative research to research focused on development (in US government terminology 6.1, 6.2, and 6.3 level research). In every research case, different burdens fall on researchers, many of which could be profitably automated with the correct computer tools. For example, it is very tedious to maintain laboratory record in the search for new drugs: software tools are needed to maintain all past and present laboratory records. A third dimension of burden is the effects of teamwork on research and the complexity introduced by collaborative and cross-disciplinary efforts.

The approach to understanding the research process was to interview researchers.

Researchers

The purposes of the interviews was to find out what types of tools each researcher used, as well as how effective the tools were. Questions were asked about what they researched, how they manage their research and the tools they use. The interviews were not extremely formal, so the questions could be adapted during the interview process. The first interview helped in setting a format for future interviews and in creating valuable questions for future interviews. Below is a general format of the interview questions followed by a summary of each interview.

Questions

This general format of questioning was followed for the interviews but maintained the interviews in a more conversation-like arrangement. Throughout the interviews, researchers were able to expand on their ideas as well as ask questions, which in turn, generated more questions.

1. Why types of research do you do?
   a. What do you consider when planning research?
2. What are some of the tools you use for your research?
   a. Where do you obtain your tools for research?
   b. How timely are the tools you are using?
   c. What tools do you use for proposal preparation?
      i. Reports
      ii. Budgets
      iii. Regulations
      iv. Formats
3. Do you prefer using primary sources, secondary sources, surveys, interviews or observations? Why?
   a. How do you know your sources are credible?
   b. Do you know many different sources you need to use in order to sufficiently cover your topic? If so, how?

Dr. Gabe Garcia, Mechanical Engineering

Dr. Gabe Garcia of New Mexico State University's Mechanical Engineering Department is currently researching signals using pattern recognition. More specifically "eddy current (magnetic field) data from a steam generator from a nuclear power plant. 'Any time a magnetic field is not uniform I will be able to pick it up. This will enable damage to be detected'." The interview started about some tools he used for research. Dr. Garcia stated that he used programming tools for his research. He stated that since his research consisted of many data points and plots, he used Matlab for calculations. Matlab was ideal for him.

"Mathworks program Matlab is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include Math and computation algorithm development, data acquisition, modeling, simulation and prototyping data analysis, exploration and visualization, scientific and engineering graphics, and application development including graphical user interface building." (Mathworks, 2005)

Because of all these capabilities, Dr. Garcia can input data into Matlab and can further analyze specific patterns. Matlab helps him perform complex mathematical operations. He states that getting the program to do these operations is pretty complex, since Matlab is also very particular in keyboard characters. It was stressed that using Matlab is "very time consuming." But once the programming is done, it is very easy to read and understand what is happening from the data.

He was also questioned on the types of tools he use for maintaining reports. He stated that he had to do reports about every four months and some were yearly. He simply used the basic Microsoft Word for his first report and then edited this first report to serve as the second report and so on. When asked what tools he used for proposal writing, he again stated that he used the basic Microsoft Word if he needed to. Usually funding agencies like the National Science Foundation require a certain format for a proposal to be accepted. These formats are available on their website.

We asked about collaborative projects during our interview, but Dr. Garcia mentioned that most of the collaborative projects he had done, he had just worked on the proposal part of the project.

Although graduate students are not tools, Dr. Garcia pointed out some positive and negative points of having a graduate student assist with research. One positive point was that the graduate student would be able to work on the research when he was not available, they would be able to perform tasks while he was busy with something else. But on the other hand, if he were to receive a graduate student who was not a hard-worker at all, time, money, and effort would be wasted.

Dr. Doug Gillan, Psychology

Dr. Doug Gillan is the department head of the Psychology Department at New Mexico State University. The interview started out by giving Dr. Gillan a brief background of what it was that we were researching in and how it tied in to interviewing him. As a researcher in psychology with emphasis on applied, cognitive, and perceptual psychology for the last 15 years, he stated he thought our project would be interesting but along the lines of impossible. A lot of the research he does deals with how people read information of a display. Information being graphics, equations, or a table of data, this is a large amount of his research. His research also consists of cognitive and perceptual problems that are faced when controlling multiple robotic controls.

The next topic of discussion was what Dr. Gillan considered when he planned his research. He stated that his research was not planned because often times research objectives change with results. He has planned research before but rarely makes it past the first two steps. Dr. Gillan stated that frequently "research is driven by what comes out of the studies" or the results that can change his focus.

On this subject he simply stated he's "a lot smarter after an experiment is conducted." By this Dr. Gillan means that by seeing his final results he sees where he could have changed a task to obtain more efficient results. He feels that if he planned, he would have to be smarter than he is now to plan for what he wants to happen.

When the idea of an intelligent assistant was mentioned to him he said it was unheard of. An intelligent assistant would be able to research other peoples' work to determine if his plan-of-action was going to be effective or not.

For research methods, he mentioned a research project done with a student Melanie Martin on the credibility of web or internet sources. He then suggested we speak with Melanie Martin so we could further clarify her project. He stated that his project was time-consuming in producing results. If he could have read in detail what a previous researcher had presented in a research project similar to the one he was carrying out, it would have saved him time and he could have

9

directed his research to produce more effective and useful results. This suggests pre-researching should be a step in the research process.

For the pre-researching task it is important to cite, depend or rely on credible sources. Dr. Gillian stated that in order to determine what sources he uses he makes sure that they are within a certain parameter relating to the interest of the research. In other words, they are "constrained" and "credible because they are published in a reference journal" or by "researchers [he] knows."

Dr. Vincent Choo, Mechanical Engineering

Dr. Vincent Choo is an associate professor in the Mechanical Engineering Department at New Mexico State University. His area of interest is in polymers and composite material. Because of confidentiality rules he was unable to say what he was presently researching.

Upon starting research, Dr. Choo states he does information searches on the issue being researched. He starts background research through literature searches. Background research is ideal in learning basic concepts, important issues, or key people that specialize in similar research. This is very important if a researcher is new to the subject or is participating in interdisciplinary research. This important step, as stated by Dr. Choo, can help find similar research or research on the same topic already done.

When Dr. Choo conducts a literary search he uses tools like the Engineering Index or the online version, the Compendex. This returns many pieces of literature with the subject that is related to the one entered in the search. Dr. Choo states that this still is not completely helpful if the article he is looking for is not available. He says filling out requests like interlibrary loans is a time consuming and slow process.

A tool Dr. Choo suggested was going on to the Internet and searching for a subject in order to return credible relevant references that had already been peer reviewed. This will ensure that the literature has been reviewed by people with specialized knowledge in a certain areas.

A search was conducted to see if there was a tool like the one Dr. Choo suggested. He had suggested something similar to the Science Citation Index. The Science Citation Index is a tool used by researchers to find sources within a specific criteria specified by the user. Then the Science Citation Index will tell how many times it has been cited in other works. If the literature is cited quite a few times then it will gain credibility. The problem with this is that if a not credible writing is cited over and over, it will gain false credibility. To ensure credibility the researcher or user would have to read through the whole work and determine whether or not they think it is a reliable source. One other recent way to limit searchers to academic sites is to use the recent Google Scholar research service. This also can check whether the researcher's institutional library has a subscription to a particular electronic journal.

Dr. Melanie Martin, Computer Science

Melanie Martin was a graduate student at New Mexico State University in the Computer Science department at the time of our interview. Her research interests are natural language processing, computational linguistics, and information retrieval. She has worked with Dr. Peter Foltz (Psychology) in applying Latent Semantic Analysis to model discourse. Currently her dissertation research is "to develop a measure of the reliability of information found on medical web pages" (Martin, 2004).

Upon talking with researchers about different tools they used when doing research, one main topic that came up repeatedly was the need to find reliable and credible sources. By sources it is meant, articles, web pages, documents and other related resources. One researcher that was interviewed stated that a tool that would help him in doing his research would be a tool that found only the articles that were peer reviewed.

Melanie Martin's current dissertation is directed toward finding the "reliability of information found on medical web pages" (Martin, 2004) but we hope to be able to use techniques, such as hers, within the Virtual Research Partner.

Researcher Tool Summary

Through the interviews it was found that instead of using project management software or task software or report building software, that using Microsoft Word, PowerPoint, Excel, the World Wide Web and the Engineering Compendex were sufficient enough for them to produce reports, proposals, posters, planning or brainstorming. When asked why current available software packages were not utilized, most of the answers stated that they were time consuming. The interviews showed that the most common tools used were the easiest tools to use.

COMMERCIAL SOFTWARE PACKAGES

As part of the study current commercial software packages that might be carrying out some of the functions needed in a VRP were studied. This is an evolving field and sometimes it is difficult to sort through the hype to find exactly how effective a package is in supporting research or business. The packages examined fell under the general headings of information discovery, content management, knowledge management, and update software.

Information Discovery

Information discovery is described as a "general term covering all strategies and methods of finding information" (Arms, 1999) or by "actively seeking out new sources of information from locations of which the user may be unaware" (Foner, 1994).

Information discovery is a central idea in research. That makes this task very vital and critical.

Because there is so much information available on all subjects, it is important to try to obtain the right information that is relevant to the subject being explored. There are many ways to obtain information. Some examples are books, search engines on the World Wide Web, scholarly journals, the Science Citation Index, and by talking to experts in specific fields. The important part of finding information is doing it effectively. By this, it is meant finding tools that are efficient in bringing credible information in a timely manner.

In the Internet Era many companies have tried to create software packages that contain several tools that assist the user in key tasks. Most of the software packages are geared for major

companies and large businesses. An example of an information discovery tool is a software package created by **Entopia**. **Entopia** uses their "K-Bus" software which is an "Infrastructure software designed to automatically and dynamically discover and deliver relevant content and experts into the day-to-day business process" (KMWorld, 2005). With Entopia's software, many different types of information useful for businesses or companies can be found in many different forms. Such as "content, experts, information sources, interactive concept maps and social network maps within various business specific applications, portal frameworks or software infrastructures." (Entopia, 2005) Entopia is able to find information with three "core components." (Entopia K-Bus, 2005) Those of which are *Intelligent Content Connectors, K-Bus Metadata Repository, and K-Bus Application Services.* The intelligent content connectors connect to "information sources" such as file shares, emails, documents, HTML content, other repositories, and so on. The K-Bus Metadata Repository "processes and indexes the information extracted from connected enterprise data sources to create metadata." Being able to tell how, when and by whom a specific set of data was collected and how it was formatted makes it easier to backtrack through information. As for Application Services, these "use the metadata do deliver never-before-known insight." These services include enterprise search, social networks mapping, content visualization and expertise location.

Enterprise search allows users to search for specified custom information of documents, expert information, and other relevant sources from with in the company's databases. Social networks mapping is able to "identify topic-based social networks to visually depict the flow of information across an organization." This means that patterns of communication, relations and dealings with in an organization are monitored and noted. This will help the software enhance information flow and help users create relationships within the organization based on information used. Content visualization is simply an easier way to navigate through large amounts of information. It is structured by a "graphical map of the key concepts within a set of content" that allows the user to easily understand and obtain." This visualization map is called the "the Entopia K-Map." Social networks mapping is able to "identify topic-based social networks to visually depict the flow of information across an organization." This means that patterns of communication, relations and dealings with in an organization are monitored and noted. This will help the software enhance information flow and help users create relationships

within the organization based on information used. Expertise Location is closely tied to social networks mapping. Expertise Location links users to other users with expert knowledge in the information area being searched. This can also be described as being able to "find the people in the organization with the most relevant knowledge or expertise." Entopia claims this helps "gain instant access to valuable resources" or their users, "drive collaboration, knowledge-sharing, innovation", "improve employee productivity and customer satisfaction." This will all be achieved by "decreasing the amount of time needed to locate the *most* appropriate person within the enterprise to answer a question."

After searching through Entopia's descriptions of their software, it is not very clear how much set-up time is needed. It is also not obvious how much updating is required of the user. Although Entopia is very helpful it needs a lot of time in order to be fully operable. Entopia has some capabilities and qualities it would need to be built-in to a fully functioning VRP.

Content Management

For the Virtual Research Partners (VRP) Research, Content Management is "ways to store, index, search, retrieve and organize…a growing collection of disparate items" (Lipton, 2004). It is also "the process of sharing information vital to an organization" (Mitchell, 2004).

Content Management is a very important area while dealing with information discovery. Mass information is discovered and retained. It is important to try to keep everything organized for easy management for later use. If you are using your information to provide research insights this management will definitely make it easer to access when you need to access.

Some tools that provide content management are tools like RedDot Solutions or Hummingbird Enterprise. Tools like these can be integrated into large businesses or companies. These tools offer web content or enterprise content management solutions which allow customers to "access, manage, analyze and collaborate around structured and unstructured business content" (Hummingbird, 2005). These solutions are extremely ideal and necessary for business, but are 0% autonomous. In order for any of these "solutions" to fully operate, a user must manage and operate the software.

14

Particularly with the RedDot Solutions you can manage web content and enterprise content. For managing web content, company or business websites are stated to be "effortless with their software." The Web Content Management offers six different modules within this one tool, each specific to web content. Some of the different modules are SmartEdit, this offers the business or company to have specific web content updated by users or employees who have expertise in that area; Asset Manager offers the ability to store "all corporate images in a central and secure location" (RedDot, 2005). Site Manager is supposed to take the place of an IT manager. It will enable you to "ensure that your site is cohesive and easy to navigate."

With Site Manager, you can let different people manage a specific area on the company website, store all corporate images in a primary secure location, create page templates, edit content, translate your site to different languages, ensure your visitors can access your website and so on.

The enterprise content management includes the "Web Content Manager, with a Document Manager, Collaboration Manager, and Business Process Manager." (RedDot, 2005) All of which are extremely ideal for a business or corporation.

For Hummingbird Solutions, they offer "enterprise content management (ECM) solutions allow customers to manage the entire lifecycle of enterprise content from creation to disposition." (Hummingbird, 2005) Hummingbird, which is similar to RedDot, offers content/document, e-mail, records, and knowledge management. Hummingbird also offers instant messaging, mobility, query and reporting, data integration, and portal framework. With all these tools and capabilities there still is minimal autonomy.

Knowledge Management

Often times Knowledge Management (KM) and Content Management (CM) are treated as if no differences between the two exist. But, knowledge management can be defined as the knowledge of the organization, structure and importance of the content that is being managed. Whereas content management is the organization and structure of content (ex: files, images, documents, and other unrelated items). One way Knowledge Management can be defined is as a

"process through which organizations generate value from their intellectual and knowledge-based assets. Most often, generating value from such assets involves sharing them among employees, departments and even with other companies in an effort to devise best practices" (Santosus & Surmacz, 2001). Knowledge Management would be helpful to the Virtual Research Partner (VRP), since organization of everything related to the research enterprise would be available.

In order to get the most benefit from a company, "KM practitioners maintain that knowledge must be shared and serve as the foundation for collaboration" (Santosus & Surmacz, 2001).

Update Software

Cymfony, Inc. develops information discovery solutions for Enterprise, Internet and Wireless environments. Cymfony, based in Buffalo, NY, has extensive experience in linguistics, information extraction and natural language processing. Cymfony's flagship product, Dashboard, enables users to ask natural language questions and receive immediate answers. InfoXtract mines information from a broad array of structured and unstructured data sources and live data feeds. Cymfony's Brand Dashboard™ is the first product in a suite of business intelligence solutions to leverage the InfoXtract engine for marketing, PR and branding professionals.

Cymfony is a real-time software solution. It is targeted to businesses who are interested in tracking and keeping reports on the influence of media exposure on brands, companies, key people and messages. This software is exceedingly ideal for very competitive businesses. Through this they can determine what means of advertisement is better exposed through media.

This is also ideal for campaigning. Through Cymfony Dashboard you can track favorability of a competitor, whether or not the competitor's message is stronger or more positive than yours. Cymfony software looks useful to people who are in the competitive business. It is not really geared to help research in the technical area. The software mainly reports and tracks media exposure on what the user selects. The user can select brands, companies, key people, or messages to track. "Dashboard software is the foundation for automating the aggregation of

multiple content sources and organizing clippings to put content into context" (Cymfony, 2005). This software is autonomous but will never automatically search new topics until told to.

Software Package Tool Summary

There are many different software packages that are available. After reviewing the software packages mentioned above, it was found that with much setup time, minimal autonomy was incorporated into these software packages. As anticipated, a lot of information was needed to establish a beginning point for the software to work.

Also, these software packages seemed to be more ideal for big corporations and businesses. They did not seem to have any importance or assistance for academic research. However, some of the functionality provided by these packages would be useful to research enterprises, particularly large ones such as FFRDCs (federally funded research and development centers). Even within a university it is possible to have researchers with common interests who do not know of each other's existence.

COMPUTERS AS ASSISTANTS

The VRP is to be an assistant that would be able to autonomously research or assist in the scientific process. Many different researchers have investigated this idea of a computer as a smart assistant. Peter Hoschka describes some of the research in his book, Computers as Assistants: A New Generation of Support Systems (1996). This book is a compilation of many different projects researching computers as assistants. This book presents initial ideas of what needs to be done to have this type of assistant. Although it is clearly stated in the introduction and overview of the book that the assistant "should *not* automate tasks completely" the research goal of the VRP proposes the complete opposite (Hoschka, 1996). Hoschka states "the basic paradigm is that of assistance" and "in many fields of application the problems are either too complex or simply too numerous for any attempt to develop a machine with complete problem solving competence to succeed." Throughout this introduction, the author stresses that their

intelligent assistant is just that, an assistant. They are not trying to build "a duplicate of a human assistant", but identify the properties of a good assistant (Hoschka, 1996).

According to Hoschka the properties an assistant required was broken down into four areas of interest of Systems with Assistance Capabilities, Domain Competence, Cooperation Support, and Methodological and Tool Projects. Each of these chapters is then further divided into properties of the chapter explained by either, research projects or colleagues of Peter Hoschka.

In Systems with Assistance Capabilities, the research project REFLECT, (partially funded by the ESPRIT Basic Research Programme of the Commission of the European Communities), explored how "knowledge base systems as experts could be turned into competent problem solvers." (Hoschka, 1996) Again this section stresses that the assistant should not try to solve problems that are impossible.

In this project they proposed to improve a knowledge-based system with "suitable competence specialists" or "independent generic modules, each devoted to a special type of competence improvement." Their defined problem was to "lay foundations for future knowledge-based systems to know more about the limits of their own competence" (Hoschka, 1996). Contact with some of these people[1] was planned in order to see what the foundations for the knowledge-based systems were, so we could possibly be incorporated into the proposed VRP.

What the researchers of REFLECT started out with, was choosing an "assignment problem solver" titled OFFICE-PLAN. They showed the incompetence of the system and what needed to be done to fix it. Within REFLECT they were able to detect what the ineffectiveness of OFFICE-PLAN was (Complexity, Inconsistency, Irrelevance, Under-specification, Over-specification, Uncertainty, Errors). With these properties of incompetence the REFLECT team proposes to bring in another "problem solver" to analyze, and examine the incompetence of the primary system. With the results and conclusion of the REFLECT project, the team noted that they recognized their approach to this problem as a possible "evolutionary software development

---

[1] So far contact has not been established.

18

method going beyond the actual project objectives" (Hoschka, 1996). This easily ties into the proposed architecture of our Virtual Research Partner.

A research program mentioned in Hoschka's <u>Computers as Assistants: A New Generation of Support Systems</u> titled "Assisting Computer" (AC) was managed at GMD's Institute of Applied Information Technology (FIT). This research project focused on *"domain competence"* and *"learning and adaptively."*

Katharina Morik, a researcher in the program, stated knowledge acquisition as very difficult, and unorganized or "sloppy". She proposed three "cooperation styles for interaction" of human and assistant computer (AC) (Morik et al, 1993). For the assisting computer, one-shot learning, interactive learning, and balanced interaction were the "Balanced Cooperative Modeling." The GMD researchers then briefly described "specific properties that must be met by an assistant system for the construction of domain models" (Wrobel, 1988). Those properties are Flexibility, Reversibility, Integrity and Consistency Maintenance, Liveliness, and Inspectability

They then present the "MOBAL system". This is a "multi-strategy learning system that consists of a collection of cooperating learning modules organized around a knowledge representation subsystem" (Michalski, 1993). This stressed how the capabilities of MOBAL can be "embedded into other application systems" (Hoschka, 1996).

"Endowing a Virtual Assistant with Intelligence: A Multi-Paradigm Approach" is a project of the Intelligent Systems Research Group 1998 Universidad Politecnica de Madrid, UPM. This research, presented at the AMEC SIG meeting, February 4th 2003 in Barcelona, is another study of computers as assistants. It was presented by Josefa Z. Hernandez and Ana Garcia Serrano from the Department of Artificial Intelligence at Technical University of Madrid (UPM), Spain. Their research project proposed a "virtual assistant on risk management" (Hernandez & Serano, 2003). The main functions of this assistant are to "listen, understand and respond in context of conversation." (Hernandez & Serano, 2003) This was proposed to be achieved by speech recognition and synthesis, using knowledge based assistant, case-based reasoning, and

information retrieval support. The conversation role would be supported by retrieval of "current of past user-system dialogues" (Hernandez & Serano, 2003).

In their presentation they compare two assistants, or assistant tools. The present assistant is "Winterthur's Online Assistant," which is in German, and the VIP-Advisor vision. (Hernandez & Serano, 2003). The present edition only has visualization and risk management assistance whereas the vision has visualization, online translation, speech synthesis, speech recognition, dialogue-based interaction, natural language analyzer, natural language generator, case-based reasoning, and risk management assistance.

There are two main flow charts that tell the way the VIP-Advisor vision is supposed to flow. Referring to the web addresses, attached to the references (see references) will help in understanding the remainder of this discussion.

The natural language processing (NL processing) is planned to understand what the speaker is trying to say as in human-human interaction. The main task is to "retrieve not only the meaning, but also the intention of the speaker (and even the emotional state)" (Hernandez & Serrano, 2003). They want the NL processing to "codify" using "semantic structures" and for generation of natural language will be in the form of sentence templates which are adaptable.

For maintaining what is going on in the conversation, they have designed an Interaction agent. This agent has several main tasks which are mentioned in the presentation as "managing the evolution of the conversation in a coherent way." (Hernandez & Serrano, 2003) This involves "keeping track" of the conversation and decision making in what to do next.

Second task is to "ask the intelligent agent for information when the answer to the user requires it," and finally to "deliver the answer together with indication on how to provide it. This agent works off of incoming communicative acts and in turn produces results with communicative acts. In order for the above-mentioned agents to work, the VIP-Advisor needs to request information from either the user (ex: Are you ready to begin?) or from the case-based reasoner (CBR) and the Risk Manager.

ARTIFICIAL INTELLIGENCE FUNDAMENTALS RELEVANT TO THE VRP

For the VRP to function two key capabilities are needed. The first is the ability to represent and manipulate knowledge; the second is a control mechanism that will support interaction with the researcher. Many other capabilities would improve the interaction, for example good quality language generation, speech understanding, and image understanding. All these are gradually developing fields and in limited domains (such as a restricted research field) may function reasonably well. This study concentrated on knowledge and control.

Representing Knowledge

Ontological Semantics (Nirenburg and Raskin, 2005) is a *knowledge representation language* that includes three interrelated sublanguages for representing text meaning (TMR or Text Meaning Representation), for representing conceptual models of the world (i.e., ontologies, including script-like structures and inference rules, and "fact repositories" or FRs) as well as for representing knowledge about natural languages. The resources encoded in these languages presently include an *ontology* of about 6,500 concepts (or over 100,000 knowledge elements). These resources are used to support inferencing as needed for such cognitive tasks as understanding and responding to queries. By extending the ontology and its associated lexicon to a particular research domain we can provide support for a system to use knowledge about the domain and to interact with the user. The actual construction of such knowledge is a complex and labor-intensive task. We give a small example here of how the ontology can be used to represent basic information on chemistry.

Since the TMR has been developed to represent the meanings of texts, both queries and responses may be readily represented. For instance, the query:
Which element is classified as a noble gas at STP?

(1) hydrogen   *(3) neon*

(2) oxygen   *(4) nitrogen*

would be represented in TMR as follows:

      chemical-element

```
name            X
standard-state  gas
atomic-number         or (2 10 18 36 54 86 118)
```

and the correct response,

Neon is classified as a noble gas.

would be represented as:

```
chemical-element
name          neon
stable-state  gas
atomic-number 10
```

Here, *which element* is treated as a request for an element name, *noble gas* as a particular specification of the properties stable-state and atomic-number and *is classified as* as essentially "has the properties of." What is worthy of note, however, is that all these correspondences between text and representation need to be expressed in the Ontological Semantics.

The ontology contains conceptual knowledge of the general types of objects, events and properties in the world. Ontological concepts have a name, at least one supertype from which it may inherit properties, possibly one or more subtypes, and a set of properties. In the case of object concepts such as chemical-element, the relevant properties might include symbol, atomic-number, atomic-weight, standard-state, color, melting-point, boiling-point, etc. Event concepts have in addition a special set of relationships or roles (e.g., agent, theme, or instrument) and may be composed of an interrelated set of sub-events. Such complex events are the mechanism for including inference-supporting causal chains and scripts. Property concepts are made up of one place attributes (e.g. color, temperature, or standard-state) and two-place relations having a domain and range (e.g., contains, measured-in, measured-by).

The Fact Repository contains instances of objects, events or properties that the system has encountered before. The ontology and the FR are structurally one and the same, related to each other through the use of the instance-of relation.

Blackboard Systems

The normal method for building a computer system is as a chain of modules and decision points which communicate through passing data from one module to the next. The system may also contain internal data (for example a database) which may also be modified during program execution. The system most often is deterministic, given a particular start configuration and input data the system will pass through the same states and produce the same outputs. Non-deterministic behavior can also occur. In the case of a simulation, using a probabilistic model for event occurrence, the results will change each time the simulation is run. This allows an exploration of different potential outcomes produced by the model of a situation. Some parallel computer systems can also inadvertently produce different results depending on the exact timing of different streams of execution in the system. Non-deterministic execution is a desirable property in systems intended to display "intelligent" or at least interesting behavior. If a system is designed to interact with a human user in a humanlike manner then its having varying outputs in the same or similar situations makes it seem less computer like. Non-deterministic approaches can also be used to prevent a system being stuck in some corner of a problem space. Non-deterministic behavior does make it difficult to debug and test a system and some way of allowing this to be switched-off for debugging is usually needed. In simulations this is done by always starting the pseudo-random number generator with the same seed number.

A system of the type that is envisioned for the VRP may appear non-deterministic even when it is behaving deterministically. The amount and complexity of the data that the system is using and the fact that it needs to remember its previous behavior over a period of time mean that its response to input from the user will vary from one occasion to another. This leads to the other problem in designing a computer system for such a complex task: just what should the order of execution, the structure of a program be? One attractive way to avoid, or at least postpone, this problem is to use a form of "blackboard system." The analogy is straightforward. A bunch of people, experts in various subjects, stand around a blackboard and solve a problem by taking turns at adding their expertise to the evolving solution which is written on the blackboard. The computer equivalent of this approach is shown in figure 1 below. A data area, the Blackboard, is accessed by a set of knowledge sources (software modules). The knowledge sources (KS) themselves can have access to other data and are all permitted to read and write information on

the blackboard. In addition some control mechanism needs to be implemented to ensure that one or more knowledge sources does not hog the blackboard to the exclusion of the others.

"The blackboard approach provides freedom from message-passing constraints. The message-passing paradigm, although modular, requires a recipient of the message as well as a sender. Often the recipient is not known or the recipient might have been deleted. In the blackboard approach, the "message" is placed on the blackboard, and the developer of the module is freed from worrying about other modules" (Nii, 1986). The data is then "posted" on the blackboard for specialized subsystems to review and formulate estimated solution(s). This is ideal for questions without solutions or an extremely complex solution(s). "The blackboard approach was designed as a means for dealing with ill-defined, complex applications" or solutions (Corkill, 1991).
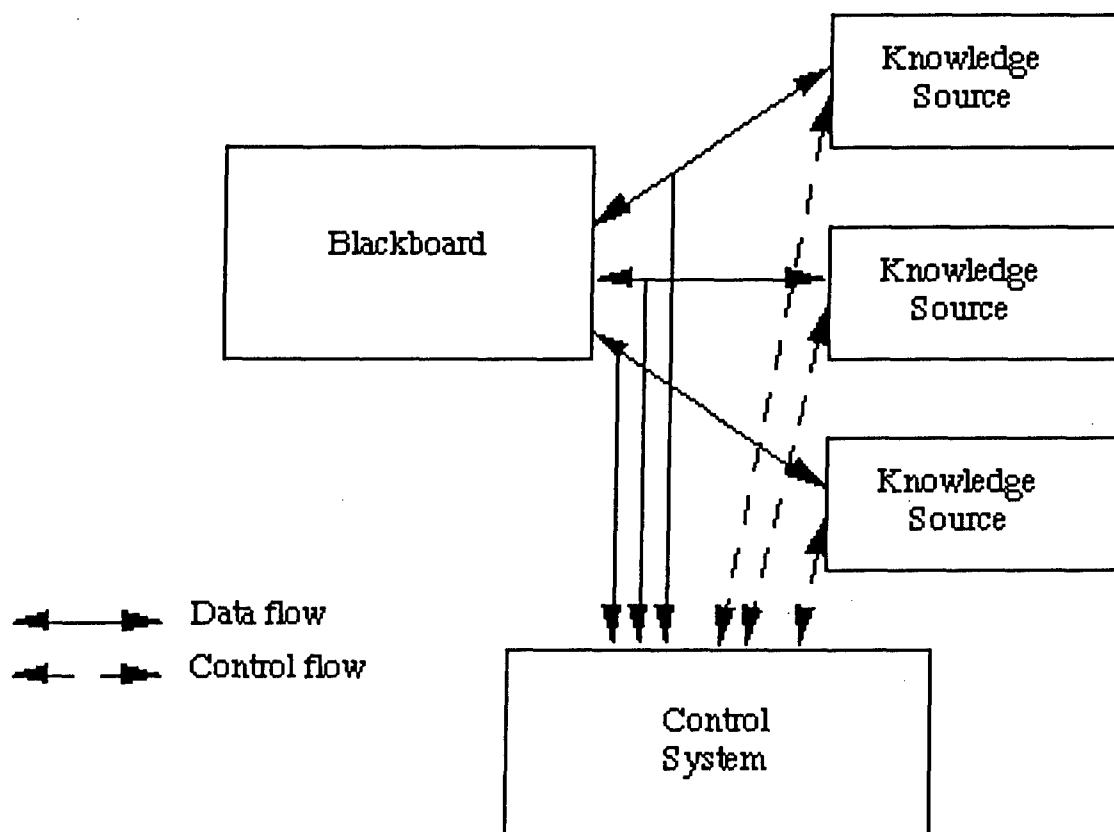
Figure 1. Generic Blackboard System

Choosing an appropriate control mechanism is one of the key problems in designing a system based on the blackboard model. Simple mechanisms such as round robin allocation may produce a system that is painfully slow, especially when the number of knowledge sources is large. Ideally the control mechanism should schedule sources that have a good chance of responding to new items appearing on the blackboard. For example a KS might post a request for information and request that it is only scheduled when some other KS has provided that information. The control system then maintains a list of triggers and associated KSs which should be scheduled when a trigger is activated. Too rigid an implementation of this strategy can lead to a state where everything grinds to a halt and methods for avoiding deadlocks need to be implemented.

Work on Blackboard Systems and Available BB Platforms

"Beginning with Penny Nii's AGE skeletal blackboard framework that was developed at Stanford University from 1977 - 1982, academic researchers have built tools for their own blackboard system research. Most notable of recent academic research tools are Barbara Hayes-Roth's BB1 system (which can be licensed from Stanford) and University of Massachusetts Amherst GBB framework. These systems have the advantages of low-cost and complete source code" (Corkill, 1991).

"They have the disadvantages of limited documentation and supporting appropriate KSs like determining the structure of the blackboard and the objects needed, selecting a control approach, determining control knowledge, etc., all must be determined when developing an application. For someone developing a blackboard application for the first time, these choices may be intimidating. However, for an experienced blackboard application developer, these same choices present opportunities for tailoring a high-performance approach to the problem. For the novice developer, the flexibility of the blackboard architecture allows an incremental approach to complex problems" (Corkhill, 1991).

A DESIGN FOR THE VRP

Based on investigations so far the NMSU team proposes the design discussed below as a basis for implementing the first level of an electronic assistant to support one or more researchers. The VRP should be adaptable to the needs and preferences of a particular researcher and field and will use a

25

society of intelligent software agents that will understand the researcher's goals and assist him or her in attaining them. The initial version of the VRP is intended to provide help to the researcher in the areas of data mining, document retrieval and proposal and presentation development. The plan is to develop agent-based computational architecture using a software base developed at Stanford (Stanford, 2003), which will be creatively adapted to the needs of VRP.

The idea is to develop and test increasingly sophisticated models of research activity. These will be addressed to the needs of specific researchers. VRP will carry out, on its own initiative, support activities for the researcher and provide him or her with an unobtrusive sophisticated interface that will help the researcher in the following types of facilities: a) **enabling capabilities**, including automatic pre-processing of material in foreign languages and/or filtering of otherwise overwhelming text volume; b) **completion capabilities,** by continue to pursue a line of work (e.g., finding information about a particular topic) that the researcher has suspended due to time constraints; this will include collating information from different sources and media and c) **memory-aid capabilities**, including the provision of active checklists to track the progress on a task or overall progress on a project. This capability will support the researcher by recognizing situations in which work related to collaboration is mandated. In this area new standards and recommendations may be appearing on a regular basis.

The work will include the following major tasks:

1. **modeling researchers and the research process** and interpreting the results in terms of an internal ontological representation of a detailed hierarchy of researchers' typical goals and strategies (plans) that the researchers typically follow to attain those goals; this task will also prove that ontological work in the style of the OntoSem (Nirenburg and Raskin, 2005) approach at NMSU strongly facilitates encoding of "fuzzy" and "contradictory" knowledge to capture a psychological model ;

2. **acquiring actual prior and tacit knowledge** about a domain or domains, by incorporating it in the ontology and its corresponding fact database; this task also includes massive intake and organization of data from text sources into the fact database through an extended, ontology-informed use of software agents such as IR, IE, MT, text summarization, goal- and plan-directed reasoning and hypothesis formation;

3. generalizing and extending the goal and plan hierarchy obtained in Task 1 by **generating novel strategies (plans)**; this task will also experimentally prove that research on logics, reasoning and theorem proving methods only becomes realistically applicable when experiments in it are conducted on the basis of a realistic-coverage and adequately fine-grained ontological and fact knowledge base and in the framework of an actual analytic task;

4. developing a **control** system for triggering the software agents with the goal of performing the "complementary" objectives specified by the goals and plans that the researcher chooses but has not completed, with the objective of reminding the analyst of lacunae in analysis as well as anticipating the analyst's needs (NIMD Areas 3 and 5)

5. developing and integrating **human information interaction** methods for interacting with the researcher.

VRP IMPLEMENTATION

Human Information Interaction

One problem that has to be addressed to support human information interaction is the disjoint between the resources used by a human to keep track of his/her archive of material and the resources needed by a machine to keep track of the same material. The researcher's traditional boxes of notes and reprints and his/her personal knowledge and experience are still a good model of how human information interaction proceeds, shuffling a variety of constraints and possibilities both physically, on the desktop or whiteboard, and mentally. The internal parallel data for manipulation by the computer, in our case, is the Ontosem ontology and the associated network of facts held in the fact database. What is needed for the VRP is to provide a seamless mapping between machine and user representations. We intend to provide a significant degree of support for a variety of analogies which allow the researcher to manipulate the information using interfaces which support a researcher's view of information.

A significant amount of effort will be devoted to utilizing the standard representations that are used to catalog information obejcts which can be thought of as lying midway in the continuum between human useable and machine useable.

Idealized VRP/User Interaction

The intelligent system must be able to do, at least, the following: perceive, understand, predict, manipulate, learn, remember, reason, and have rationality, to name a few. In "Artificial Intelligence: A Modern Approach" four types of intelligent systems are mentioned: systems that *think* like humans, systems that *act* like humans, systems that *think* rationally, or systems that *act* rationally (Russell & Norvig, 2003). We would like our assistant to exhibit all four types of traits.

If one is building an intelligent assistant, the system must be a rational system, since it is defined as one that does the right thing. Autonomy must be decided. A system with no autonomy is achievable. A system with autonomy is much more complex.

An idealized user interaction between researchers and the Virtual Research Partner (VRP) will be initiated by the user with text, speech, images, formulas, graphs or data of those types. This implies the VRP will be able to recognize all those forms of data and be able to process the meaning and content of each. Given that many of the capabilities needed for a VRP are complex research questions some way is needed to allow incremental developments to be added to the VRP as they become available. This implies that we must allow insertion of modules of which we can not currently specify in detail. Our proposed approach is to use a blackboard (BB) architecture. Then data stored in the system is associated with records containing fields indicating its type, source, and relation to other data. These data records can also contain requests for further processing.

Blackboard agents will operate on these data elements and add further elements. Through the blackboard, data is posted and allowed to be interpreted by knowledge sources. The VRP knowledge sources will understand and acknowledge the meaning and content on the Blackboard so that relations between future, current and past data can be made.

To produce even a basic VRP there will have to be many knowledge sources operating on the Blackboard. Some will be metatasks such as KS scheduling, system history, and user interaction. Others will provide generic components needed by many research efforts graph

production, background research. Finally a really large scale KS can be built out of the generic components: proposal preparation, project coordination, research paper development.

Interactions with the VRP will be asynchronous. If the researcher has initiated communication with the VRP and begins having difficulties proceeding, the VRP, without being told, will suggest relevant tasks to stimulate new directions for the researcher to continue in.

The knowledge sources (KS) mentioned will rely on specific generic knowledge sources (GKS) to help with specific tasks. For Proposal Preparation knowledge sources, GKS's will handle technical and contract parts of a proposal independently. In the contract proposal preparation KS other KS such as analyzing the call for proposals, producing budgets, and satisfying funding regulations will be needed. The proposal packet developed to answer a request for proposals must satisfy funding regulations and also must meet the requirements and guidelines of the university. The knowledge source for proposal preparation needs, under the direction of the researcher, to be able to complete these tasks. If the knowledge source needed help from other sources the issue would be placed on the blackboard system until a solution could be sent back to the proposal knowledge source from another knowledge source. If the issue needed to be answered by the researcher, a question would be formulated, posted on the blackboard and sent to the researcher and then submitted back to the proposal knowledge sources (See Blackboard Figure 1).

Applications and Components

These agents described below are an outline of what capabilities VRP should have.

1. Information Discovery: The information discovery agent will be "actively seeking out new sources of information from locations of which the user may be unaware" (Foner, 1994).

2. Knowledge Management: The knowledge management agent will organize structure and know the meaning of the content that is being managing.

3. Content Management: The content management agent will store, index, search, retrieve and organize all content such as files, images, documents, and other unrelated items.

4. Proposal Writing: The proposal writing agent will be able to write proposals for the researcher. As specified by the researcher the agent will know which format to use. Proposals submitted to DARPA, NSF, or NASA all have specific formats that need to be followed.

5. Scheduler/Task Writer: The scheduler/task writer agent will keep track of the postings on the blackboard. For instance, if a posting has been on the blackboard for a while and has not been answered or acknowledged, the agent will inform the user it has been inactive. Then from that response, the agent will either file it in the database/history, or continue to find a response to the posting.

6. PowerPoint/Poster: The PowerPoint/Poster agent will be able to create a PowerPoint presentation or a poster of the report or research. This will be for presentations, workshops or conferences.

7. Questioner: If at anytime during the research or project, the researcher gets a writer's block, or runs out of ideas, the questioner agent will question the researcher in order to try to generate ideas. Ex: "How about____" or "What if____".

8. Background Research The Background agent will be able to find previous works of research that has been conducted that has relevance to the project the researcher is working on. That way the researcher can start from where someone else left off, with out having to repeat research that has already been done.

CONCLUSIONS AND A PROPOSED APPROACH TO FUTURE WORK

No one is attempting any effort of the scope needed for a true VRP. The group in Spain is working on many of the interaction techniques that would be needed by a VRP, but their domain topic "risk analysis" is a very limited one.

The researchers interviewed used a very limited set of tools to support their work. One thing that emerged strongly was the need for better ways of searching and mining the literature. In part this is a problem created by the booming publishing and conference "industries". Conferences now have many attached workshops and there are more and more of them. Journals appear in many languages and an English centered approach to research is perhaps becoming less tenable in some fields. (One example is the field of parasitology; little covered in US medical journals, but a major topic of interest for journals in South America). This is one area in which a VRP could be constructed using current information retrieval and extraction technology, with search methods tuned to academic materials and sources. The other areas of proposal preparation and submission and preparation of parts of the publishing process may also be amenable to at least partial automation.

The problem here is the amount of knowledge encoding work that would need to be done to support these efforts. This is truly a Herculanean task. A potential approach round this would be to recruit a community of vounteer workers to build components to fit into a general framework. The model here would be similar to that which allowed the development of Linux and also the Wikipedia. In these cases a basic infrastructure existed which was added to by a community effort.

This approach is currently being investigated by implementing a new application program interface for interacting with the Stanford Blackboard Kernel System (BBK). This will allow knowledge sources in a variety of languages to interact with the VRP blackboard. If further resources were available we could then extend the ontology and lexicon to describe the tasks of proposal preparation and one small research area. An ontology of blackboard metatags (descriptions of blackboard data types and actions) is also needed. An example "mini" VRP

system could then be assembled and made available as a prototype for groups wishing to add further capabilities and modules. This bootstrapping approach would seem to be one way of involving multiple communities in designing both knowledge representations and knowledge manipulations. The basic requirements would be a distribution and integration mechanism and clear instructions on how to build different component types adressed to non-CS researchers.

REFERENCES

Air Force Research Laboratory. (2005) AFRL Database – http://www.biodyn.wpafb.af.mil

Arms, W. (1999). From the 1999 manuscript of *Digital Libraries*, by William Arms, (c) 2000
M.I.T. Press. Available: http://www.cs.cornell.edu/wya/DigLib/MS1999/glossary.html

Chalupsky, Hans. (2000) "OntoMorph: A Translation System for Symbolic Knowledge". In
*Proceedings of the Seventh International Conference on Principles of Knowledge Representation
and Reasoning (KR-2000)*, Breckenridge, CO, April 2000

Colby, Kenneth. (1973). "Simulation of belief systems". In R. Schank and K. Colby (eds.).
Computer models of thought and language, 251-286. San Francisco: Freeman.

Corkill, D. (1991). "Blackboard Systems". *AI Expert, Vol: 6, Num: 9* Available:
http://mas.cs.umass.edu/pub/paper_detail.php/218

Cowie, J. and W. Lehnert. (1996). "Information Extraction", *Communications of the ACM
special issue on Natural Language Processing*, January.

Cymfony. (2005). *Cymphony Dashboard.* Available: http://www.cymfony.com/sol_what.asp

Entopia. (August, 2004). *Solutions for Information Discovery.* Available:
http://www.entopia.com

Foner,L. (1994). *Glossary of Agent Community Terms,*
Available:http://foner.www.media.mit.edu/people/foner/Yenta/glossary.html

Hernandez, J. & Serano A. (2003). *Endowing A Virtual Assistant with Intelligence: A Multi-
Paradigm Approach*. Department of Artificial Intelligence Technical University of Madrid
(UPM), Spain. Available: http://www.cs.uu.nl/~dignum/hernandez.pdf

Hoschka, P. (Ed.) (1996). *Computers as Assistants: A New Generation of Support Systems*. New
Jersey: Lawrence Erlbaum Associates

Hummingbird. (2005).*Enterprise Content Management Platform*. Available:
http://www.hummingbird.com

KMWorld. (August, 2004). *Trend-Setting Products*.
Available:http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=9575

Knox, T. (2004). "Intelligent Systems as Virtual Research Partners" in *AFRL BAA 04-05-HE*

Lenat, D.B. and R.V. Guha. 1990. *Building Large Knowledge-Based Systems*. Reading, MA:
Addison-Wesley.

Lipton, R. (2004). *What is Content Management?*
Available:http://radio.weblogs.com/0100059/stories/2002/04/09/whatIsContentManagement.ht
ml

Martin, M. (2004). *Melanie J. Martin.* Available:
http://www.cs.nmsu.edu/~mmartin/research.html

Mathworks. (2005). *Matlab.* Available: http://www.mathworks.com/products/pfo/

Michalski, R. S. (1993). Inferential Theory of Learning as a Conceptual Basis for Multistrategy
Learning. *Machine Learning* 11(2-3): 111-151.

Mitchell, B. (2004). *What is Content Management, Really?*
Available:http://www.widez.info/general/content-management.php

Morik, K., Wrobel, S., Kietz, J., and Emde, W. (1993). *Knowledge Acqusition and Machine
Learning – Theory, Methods and Applications.* London: Academic Press.

National Science Foundation. (1996).*The Future of Science and Technology in New England:
Trends and Indicators.* Available:
http://www.aaas.org/spp/cstc/pne/pubs/regrep/NE/Nwnglnd.htm

Nii, H.P. (1986). "Blackboard Systems". *AI Magazine, Vol 7, No. 2,* p38-53.

RedDot. (2005). *CMS: Web Content Made Simple.*
Available:http://www.reddot.com/text/679.asp

Russell, S. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach. (2nd ed).* Prentice
Hall

Sandia National Laboratories. (2005). *VXInsight.* Available:
http://www.cs.sandia.gov/projects/VxInsight.html

Santosus, M., Surmacz, J. (2001). *The ABC's of Knowledge Management.* Retrieved February
22, 2005 from Knowledge Management Research Center Web
site:http://www.cio.com/research/knowledge/edit/kmabcs.html.

Stanford. (June, 2003). *Past Project: The BB1 Blackboard Control Architecture.* Available:
http://www-ksl.stanford.edu/projects/BB1/bb1.html

Weizenbaum, J. (1965).ELIZA--a computer program for the study of natural language
communication between man and machine. *Communications of the Association for Computing
Machinery 9(1):36-45.*

Wrobel, S. (1988). Automatic Representation Adjustment in an Observational Discovery
System. In: D. Sleeman, editor, *Proceedings of the 3rd European Working Session on
Learning,* p 253-262. (Glasgow, Scotland, Oct. 3-5, 1988). Pitman, London, 1988